Identity Preserving 3D Head Stylization with Multiview Score Distillation

Supplementary Material

A. LD objective

This section will go through the detailed derivation.

Recall the DDPM forward process:

$$\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}n = x_t, \quad n \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

Assume that distribution (q) of the 3D representation (θ) conditioned on generation prompt (y) is proportional to the prompt-conditioned distribution (p) of independent 2D renders (x_0^i) on different poses (i). In our setup, θ is the style-based 3D GAN layers.

$$q(\theta|y) \propto p(x_0^0, x_0^1, ..., x_0^N|y) = \prod_i^N p(x_0^i|y) \tag{2}$$

We optimize negative log-likelihood of Eq. (2) to find θ:

$$-\log q(\theta|y) = -\log \prod_i^N p(x_0^i|y) = -\sum_i^N \log p(x_0^i|y) \tag{3}$$

Define the loss $\mathcal{L}_{\text{LD}}$ as the average of infinitely many N render poses and find gradient ∇_θ to update θ via gradient descent, where π is any given pose:

$$\begin{aligned} \mathcal{L}_{\text{LD}} &= -\frac{1}{N}\lim_{N\to\infty} \log q(\theta|y) = -\mathbb{E}_\pi\{\log p(x_0^\pi|y)\} \\ \nabla_\theta \mathcal{L}_{\text{LD}} &= -\mathbb{E}_\pi\{\nabla_\theta \log p(x_0^\pi|y)\} \end{aligned} \tag{4}$$

Using Eqs. (1) and (4) and change of variables in probability:

$$p(x_0^\pi|y) = p(x_t^\pi|y)\left|\frac{\partial x_t}{\partial x_0}\right|^{-1} = \frac{p(x_t^\pi|y)}{\sqrt{\bar{\alpha}_t}} \tag{5}$$

Take the log of both sides, the partial derivative with respect to x_0^π, and decompose the right-hand side with chain rule using the relation in Eq. (1):

$$\begin{aligned} \log p(x_0^\pi|y) &= \log p(x_t^\pi|y) - \log\sqrt{\bar{\alpha}_t} \\ \frac{\partial \log p(x_0^\pi|y)}{\partial x_0^\pi} &= \frac{\partial p(x_t^\pi|y)}{\partial x_t^\pi}\frac{\partial x_t^\pi}{\partial x_0^\pi} \\ \nabla_{x_0}\log p(x_0^\pi|y) &= \nabla_{x_t}\log p(x_t^\pi|y)\frac{\partial x_t^\pi}{\partial x_0^\pi} \end{aligned} \tag{6}$$

Extend the partial gradient chain in Eq. (6) to θ from x_0^π:

$$\nabla_\theta \log p(x_0^\pi|y) = \nabla_{x_t}\log p(x_t^\pi|y)\frac{\partial x_t^\pi}{\partial x_0^\pi}\frac{\partial x_0^\pi}{\partial \theta} \tag{7}$$

where $\nabla_{x_t}\log p(x_t^\pi|y)$ is the score function estimation. Plugging Eq. (7) into Eq. (4) yields the update direction:

$$\nabla_\theta \mathcal{L}_{\text{LD}} = -\mathbb{E}_{\pi,x_t}\{\nabla_{x_t}\log p(x_t^\pi|y)\frac{\partial x_t^\pi}{\partial x_0^\pi}\frac{\partial x_0^\pi}{\partial \theta}\} \tag{8}$$

where $\frac{\partial x_t^\pi}{\partial x_0^\pi}$ is $\sqrt{\bar{\alpha}_t}$ from Eq. (1). Notice that to update θ, we do not need to back-propagate through denoising UNet and can acknowledge the UNet output as a part of the gradient. Algorithm 1 describes the domain adaptation procedure with PyTorch nomenclature:

Algorithm 1 LD with mirror and grid grads

Require: Generator $\mathbf{G}_\theta$, neural renderer $\mathbf{R}$, super-resolver $\mathbf{SR}$, depth extractor $\mathbf{D}$, depth and text-conditioned denoising UNet $\mathbf{SD}$, generator mapping truncation parameter ψ, extrinsic triplane render matrix π, mirror pose π', vertical flip operator $\mathbf{M}$, rank weighing matrix $\mathbf{W}$

for i in $\{0, 1, ..., N\}$ **do**
  $w^+ \leftarrow$ `sample_latent(`ψ`=0.8)`
  $\pi, \pi' \leftarrow$ `sample_pose()` ▷ $\mathbb{E}_\pi$
  $x_0^\pi, x_0^{\pi'} \leftarrow \mathbf{SR}(\mathbf{R}(\mathbf{G}_\theta(z), \pi, \pi'))$
  $n, t \leftarrow$ `noise_scheduler(0.70,0.96)`
  $x_t^\pi \leftarrow \sqrt{\bar{\alpha}_t}x_0^\pi + \sqrt{1-\bar{\alpha}_t}n$ ▷ $\mathbb{E}_{x_t}$
  `with no_grad():`
    $\nabla_{x_t}\log p(x_t^\pi|y)/\sqrt{1-\bar{\alpha}_t} \leftarrow \mathbf{SD}(x_t^\pi, y, t, \mathbf{D}(x_0^\pi))$
  `grad` $\leftarrow \nabla_{x_t}\log p(x_t^\pi|y)\sqrt{\bar{\alpha}_t}$
  $\mathbf{U\Sigma V}^\mathrm{T} \leftarrow \mathbf{SVD}$`(grad)`
  `grad` $\leftarrow \mathbf{UW\Sigma V}^\mathrm{T}$ ▷ rank weighing
  x_0^π`.backward(grad)` ▷ $\nabla_\theta\mathcal{L}_{\text{LD}}$
  $x_0^{\pi'}$`.backward(`$\mathbf{M}$`(grad))` ▷ mirror gradients
  `optimizer.step()`
end for

for i in $\{0, 1, ..., N\}$ **do**
  $w^+ \leftarrow$ `sample_latent(`ψ`=0.8)`
  $\{\pi\} = \pi^0, \pi^1, \pi^2, \pi^3 \leftarrow$ `sample_pose()` ▷ $\mathbb{E}_{\{\pi\}}$
  $\{x_0^\pi\}_{\text{LR}} \leftarrow$ `make_grid(`$\mathbf{R}(\mathbf{G}_\theta(w^+), \{\pi\})$`)`
  $\{x_0^\pi\} \leftarrow$ `make_grid(`$\mathbf{SR}(\mathbf{R}(\mathbf{G}_\theta(w^+), \{\pi\}))$`)`
  $n, t \leftarrow$ `noise_scheduler(0.30,0.80)`
  $\{x_t^\pi\} \leftarrow \sqrt{\bar{\alpha}_t}\{x_0^\pi\} + \sqrt{1-\bar{\alpha}_t}n$ ▷ $\mathbb{E}_{\{x_t\}}$
  `with no_grad():`
    $\nabla_{\{x_t\}}\log p(\{x_t^\pi\}|y)/\sqrt{1-\bar{\alpha}_t} \leftarrow \mathbf{SD}(\{x_t^\pi\}, y, t, \mathbf{D}(\{x_0^\pi\}))$
  `grad` $\leftarrow \nabla_{\{x_t\}}\log p(\{x_t^\pi\}|y)\sqrt{\bar{\alpha}_t}$
  $\mathbf{U\Sigma V}^\mathrm{T} \leftarrow \mathbf{SVD}$`(grad)`
  `grad` $\leftarrow \mathbf{UW\Sigma V}^\mathrm{T}$ ▷ rank weighing
  $\{x_0^\pi\}_{\text{LR}}$`.backward(grad)` ▷ grid gradients $\nabla_\theta\mathcal{L}_{\text{LD}_g}$
  `optimizer.step()`
end for

return $\mathbf{G}_\theta$

`sample_latent` utilizes the mapping network of the

generator and maps z to w^+, later to be fed to the generator. `make_grid` creates a 2×2 grid with 4 inputs. **M** is realized with `torch.flip(x,dims=[-1])`. `with no_grad()` disables PyTorch's gradient calculation. Note that each time x_0 is generated, we implicitly pass it through VAE to embed it into **SD**'s latent space.

B. Implementation details

Baselines. We train the latent mapper in StyleCLIP [40] with PanoHead's [4] generator. For StyleGAN-NADA [14] and StyleGANFusion [50], we use [50]'s official repository and modify the generator backbone to PanoHead. For [50], we utilize their EG3D config parameters for PanoHead, and implement the adaptive layer selection for [14]. For DiffusionGAN3D [29], we implement the method based on the official paper since there is no published codebase. For our baseline, we utilize our implementation of [29] with their distance loss for domain adaptation and build upon it with our proposed improvements. We stay faithful to each baseline's original hyperparameters (denoiser checkpoint selection, noise scheduler, learning rate, optimizer, etc.) unless the training diverges.

Our training parameters. We train the generator with synthetic $z_{1\times512} \sim \mathcal{N}(0, \mathbf{I})$ data for 10k iterations with batch size 1, where the truncation parameter of the generator's mapping network is $\psi = 0.8$. We use Adam optimizer with a $1e^{-4}$ learning rate. We optimize the `G.backbone.synthesis` and `G.backbone.superresolution` sub-networks of the generator `G` and freeze all convolutional layer biases, using the same configuration as [50]. The classifier-free-guidance (CFG) [20] weight and depth-conditioned ControlNet [63][1] guidance weight are set to 7.5 and 1.0, respectively. Depth ground truths are extracted from [56] since the neural renderer's depth estimations are low-resolution and require additional clipping (64×64).

As the conditional denoiser for our method and the ablation study for showing the improvements upon [29], we employ RV v5.1[2]. For qualitative and quantitative comparison among other methods, we employ the methods' suggested diffusion checkpoints in their papers and repositories[3][4].

For mirror and grid denoising, noise start timestep t is uniformly selected among `(0.70,0.96)` and `(0.30,0.80)`, respectively, where t is from $0 \rightarrow 1$. We use `DDIMScheduler` for the noise scheduler. The number of inference steps in the diffusion pipeline is always 1 since we perform score distillation.

[1] https://huggingface.co/lllyasviel/sd-controlnet-depth

[2] https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE

[3] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[4] https://huggingface.co/stabilityai/stable-diffusion-2

Figure 1. From left to right: Ours (distribution #2), ground truth unedited image (distribution #3), edited image with full-step diffusion pipeline (distribution #1).

Quantitative scores. We construct ground-truth edited image distributions using the Stable Diffusion pipeline. For the first distribution, we take images, add noise with $t = 25$, and denoise with the style prompt using each baseline's diffusion checkpoints for 50 steps, resulting in edited images. The second distribution consists of the same images stylized using domain-adapted generators. These two distributions are used to compute FID, and individual image pairs between them are used to compute CLIP similarity scores. We generate a third distribution using unedited images to evaluate identity preservation (ID) and $\Delta\mathcal{D}$. Scores for ID and $\Delta\mathcal{D}$ are then calculated between using image pairs from the second and third distributions. Fig. 1 visualizes sample images in those three distributions.

Prompts. We use empty strings for negative prompts for our method. For positive prompts, we use the following list for all methods:

- `Portrait a person in Pixar style, cute, big eyes, Disney, sharp, 8K, skin detail, best quality, realistic lighting, good-looking, uniform light, extremely detailed`
- `Portrait of a Greek statue, closeup, elegant and timeless, intricate and detailed carving, smooth marble texture, ancient Greek aesthetics`
- `A portrait of Joker from the movie The Dark Knight`
- `Charcoal pencil sketch of human face, lower third, high contrast, black and white`
- `Portrait of a werewolf`
- `Portrait of a zombie`

Rank weighing on score tensors. Fig. 2 illustrates how an SVD-based approach can decompose a stylized portrait into coarse and fine components, and then reconstruct it at different "rank" levels (k). This progressive refinement underlines how SVD can serve as a powerful control mecha-

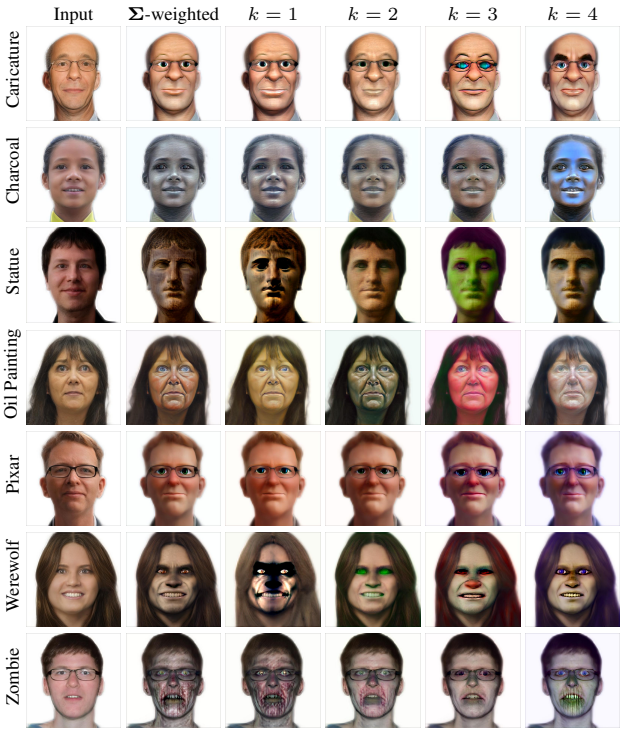

Figure 2. Demonstration of our SVD-based stylization across multiple face styles. The Input column shows the original images. Σ-weighted uses all top singular values with decreasing weights, preserving both coarse and fine features. Columns k=1 through k=4 depict rank-k approximations; as k increases, more high-frequency details are retained, resulting in sharper, more faithful stylizations.

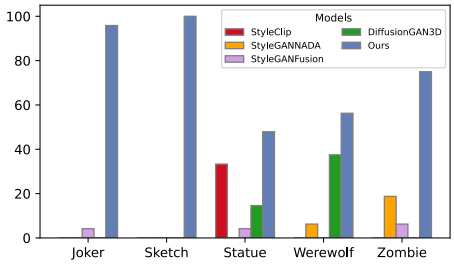

Figure 3. Percentage of user preferences. Users overwhelmingly favor ours compared to other domain adaptation methods.

nism for score distillation, letting the user dial in how many spatial "frequencies" of the style are included.

C. User study

We conduct a user study with 25 participants to evaluate the quality of 3D stylization and identity preservation across different methods. Participants are shown images generated by five different models: StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, and our own approach. For each image, they are asked to select which output best balances stylization and identity preservation. The methods are presented in random order for each image to minimize bias. Results of this user study are shown in Fig. 3. The data indicate our method is consistently preferred across all the prompts tested, with participants overwhelmingly selecting it as the best for both stylization and identity preservation compared to others.

D. Additional results

		Pixar	Joker	Werewolf	Sketch	Statue
2D	InstructPix2Pix	0.1461	0.1164	0.1427	0.0790	0.0900
	InstantID	0.0897	0.1185	0.1055	0.1218	0.1290
3D	StyleCLIP	0.1045	0.0958	0.1962	0.0878	0.1658
	StyleGAN-NADA	0.0459	0.1380	0.2617	0.0890	0.1480
	StyleGANFusion	0.1668	0.1904	0.1387	0.1168	0.1212
	DiffusionGAN3D	0.1566	0.0977	0.0922	0.1216	0.2442
	Ours	0.0326	0.0713	0.0856	0.0742	0.1122

Table 1. KID scores on the test set.

Tab. 1 reveals the KID scores on the same test set used in the main paper. Our method outperforms all baselines in KID across domains, with the exception of the Statue domain. Notably, in the Sketch stylization setting, while InstructPix2Pix reports a slightly better FID, our method achieves superior KID scores.

Figs. 4 to 9 visualizes the outputs of methods for different prompts in 360-degrees.

Figure 4. Joker edits. From top to bottom: input, StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, ours.

Figure 5. Pixar edits. From top to bottom: input, StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, ours.

Figure 6. Sketch edits. From top to bottom: input, StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, ours.

Figure 7. Werewolf edits. From top to bottom: input, StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, ours.

Figure 8. Zombie edits. From top to bottom: input, StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, ours.

Figure 9. Statue edits. From top to bottom: input, StyleCLIP, StyleGAN-NADA, StyleGANFusion, DiffusionGAN3D, ours.

References

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 4552–4562, 2023. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, CVPR, pages 4432–4441, 2019. 2

[3] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. HyperStyle: StyleGAN inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 18511–18521, 2022. 2

[4] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. PanoHead: Geometry-aware 3D full-head synthesis in 360 degrees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 20950–20959, 2023. 2, 3, 5

[5] Qingyan Bai, Zifan Shi, Yinghao Xu, Hao Ouyang, Qiuyu Wang, Ceyuan Yang, Xuan Wang, Gordon Wetzstein, Yujun Shen, and Qifeng Chen. Real-time 3d-aware portrait editing from a single image. In *Proceedings of the European Conference on Computer Vision*, ECCV, 2024. 2

[6] Ananta R Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, WACV, pages 3055–3065, 2024. 2

[7] Bahri Batuhan Bilecen, Ahmet Berke Gokmen, and Aysegul Dundar. Dual encoder gan inversion for high-fidelity 3d head reconstruction from single images. In *Advances in Neural Information Processing Systems*, NeurIPS, 2024. 2

[8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 18392–18402, 2023. 6, 8

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 16123–16133, 2022. 2

[10] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 5799–5809, 2021. 2

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 4690–4699, 2019. 6

[12] Aysegul Dundar, Jun Gao, Andrew Tao, and Bryan Catanzaro. Fine detailed texture learning for 3d meshes with generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[13] Aysegul Dundar, Jun Gao, Andrew Tao, and Bryan Catanzaro. Progressive learning of 3d reconstruction network from 2d gan data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):1–13, 2022. 1, 5, 2

[15] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, ICLR, 2022. 2

[16] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *Proceedings of the International Conference on Computer Vision*, ICCV, pages 19740–19750, 2023. 6, 7

[17] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, CVPR, pages 7498–7507, 2020. 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, 2017. 6

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, NeurIPS, pages 6840 – 6851, 2020. 3

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2021. 4, 5, 2

[21] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: zero-shot text-driven generation and animation of 3d avatars. *ACM Trans. Graph.*, 41(4), July 2022. 3

[22] Shuo Huang, Shikun Sun, Zixuan Wang, Xiaoyu Qin, Yanmin Xiong, Yuan Zhang, Pengfei Wan, Di Zhang, and Jia Jia. PlacidDreamer: Advancing harmony in text-to-3D generation, 2024. 2, 3, 4

[23] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, 2024. 5

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 4401–4410, 2019. 2

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, CVPR, pages 8110–8119, 2020. 1, 2

[26] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 14203–14213, 2023. 2

[27] Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pages 22603–22612, 2023. 2

[28] Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. Grid diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 8734–8743, 2024. 5

[29] Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. DiffusionGAN3D: Boosting text-guided 3D generation and domain adaptation by combining 3D GANs and diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 10487–10497, 2024. 2, 3, 4, 6, 7, 8

[30] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. In *Advances in Neural Information Processing Systems*, volume 36 of *NeurIPS*, 2024. 2

[31] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, ICLR, 2020. 2

[32] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pages 9298–9309, 2023. 5

[33] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations*, ICLR, 2024. 5

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, ECCV, pages 405–421, 2020. 2

[35] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 3

[36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 11453–11464, 2021. 2

[37] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, CVPR, pages 10743–10752, 2021. 2

[38] OpenAI. Dall-e 3: A system for generating images from text prompts, 2023. 2

[39] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 13503–13513, 2022. 2

[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proceedings of the International Conference on Computer Vision*, ICCV, pages 2065–2074, 2021. 1, 5, 2

[41] Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurelien Lucchi. Convolutional generation of textured 3d meshes. *Advances in Neural Information Processing Systems*, 33:870–882, 2020. 2

[42] Hamza Pehlivan, Yusuf Dalva, and Aysegul Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 1828–1837, 2023. 2

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, ICLR, 2024. 2

[44] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*, ICLR, 2023. 2, 3, 4

[45] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1 – 13, 2022. 2

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 10684–10695, 2022. 3, 5

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, NeurIPS, pages 36479–36494, 2022. 2

[48] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 18603–18613, 2022. 3

[49] Guoxian Song, Hongyi Xu, Jing Liu, Tiancheng Zhi, Yichun Shi, Jianfeng Zhang, Zihang Jiang, Jiashi Feng, Shen Sang, and Linjie Luo. Agilegan3d: Few-shot 3d portrait stylization by augmented transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 765–774, 2024. 2

[50] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris

Metaxas, and Ahmed Elgammal. StyleGAN-Fusion: Diffusion guided domain adaptation of image generators. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, WACV, pages 5441–5451, 2024. 2, 3, 4, 5, 6, 8

[51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, ICLR, 2021. 3

[52] Timothy Alexis Vass. Explaining the SDXL latent space: A complete demonstration. Available at https://huggingface.co/blog/TimothyAlexisVass/explaining-the-sdxl-latent-space, 2024. 4

[53] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 6, 8

[54] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, NeurIPS, 2023. 4

[55] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 321–331, 2023. 2

[56] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, NeurIPS, 2024. 6, 2

[57] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 7693–7702, 2022. 1

[58] Yunhan Yang, Yukun Huang, Xiaoyang Wu, Yuan-Chen Guo, Song-Hai Zhang, Hengshuang Zhao, Tong He, and Xihui Liu. DreamComposer: Controllable 3D object generation via multi-view conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pages 8111–8120, 2024. 5

[59] Ahmet Burak Yildirim, Hamza Pehlivan, Bahri Batuhan Bilecen, and Aysegul Dundar. Diverse inpainting and editing with gan inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pages 23120–23130, 2023. 2

[60] Ahmet Burak Yildirim, Hamza Pehlivan, and Aysegul Dundar. Warping the residuals for image editing with stylegan. *International Journal of Computer Vision*, pages 1–16, 2024. 2

[61] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, 2023. 2

[62] Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Ziwei Liu, and Chen Change Loy. Deformtoon3d: Deformable neural radiance fields for 3d toonification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pages 9144–9154, 2023. 2

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV, pages 3836–3847, 2023. 5, 7, 8, 2